Gaming Algorithmic Hate-Speech Detection : Stakes, Parties, and Moves

Haapoja, Jesse

2020-04-01

Haapoja , J , Laaksonen , S-M & Lampinen , A 2020 , ' Gaming Algorithmic Hate-Speech Detection : Stakes, Parties, and Moves ' , Social Media + Society , vol. 6 , no. 2 , 924778 . https://doi.org/10.1177/205

http://hdl.handle.net/10138/317300
https://doi.org/10.1177/2056305120924778

Original Article

SM+S
social media + society

Social Media + Society
April-June 2020: 1–10

DOI: 10.1177/2056305120924778
journals.sagepub.com/home/sms
SAGE

Gaming Algorithmic Hate-Speech Detection: Stakes, Parties, and Moves

Jesse Haapoja[1,2], Salla-Maaria Laaksonen[2], and Airi Lampinen[3]

Abstract

A recent strand of research considers how algorithmic systems are gamed in everyday encounters. We add to this literature with a study that uses the game metaphor to examine a project where different organizations came together to create and deploy a machine learning model to detect hate speech from political candidates' social media messages during the Finnish 2017 municipal election. Using interviews and forum discussions as our primary research material, we illustrate how the unfolding game is played out on different levels in a multi-stakeholder situation, what roles different participants have in the game, and how strategies of gaming the model revolve around controlling the information available to it. We discuss strategies that different stakeholders planned or used to resist the model, and show how the game is not only played against the model itself, but also with those who have created it and those who oppose it. Our findings illustrate that while "gaming the system" is an important part of gaming with algorithms, these games have other levels where humans play against each other, rather than against technology. We also draw attention to how deploying a hate-speech detection algorithm can be understood as an effort to not only detect but also preempt unwanted behavior.

Keywords

algorithmic systems, game metaphor, hate-speech, social media, elections

Introduction

Much of the research on algorithmic systems has focused on widely used platforms such as Facebook (e.g., Bucher, 2018), Google (e.g., Gillespie, 2017), and Uber (e.g., Lee et al., 2015; Rosenblat & Stark, 2016), with an emphasis on what platform companies do and how platform users relate to them. In this context, the notion of *gaming* has been utilized to frame situations where users try to manipulate technical systems for their own advantage (e.g., Cotter, 2018; Ziewitz, 2019). However, interactions in and around algorithmic systems can include a broader set of stakeholders and multiple levels of action beyond the user and the hosting platform. Furthermore, algorithmic systems are increasingly being deployed on various scales and in various societal contexts where the big platform companies are not necessarily central. In this article, we examine the social situation that emerged around one such system, a project that aimed to identify hate-speech from Finnish municipal election candidates' public social media postings. The project brought together different public and private organizations to develop a machine learning model (later referred to simply as *the model*) that was used to monitor candidates' social media messages for hate speech during the 2017 Finnish municipal elections. The project incurred critical responses online and even resulted in an interpellation to the government. Going beyond situations where individuals devise strategies to maximize their own gains by playing with algorithms' rules or try to outright manipulate them (e.g., Cotter, 2018), we focus on a situation where a complex set of actors is involved, and where the legitimacy of the algorithmic system is contested.

Like previous research, we mobilize the notion of a *game* to guide our analysis, but with a broader scope that aims to cover the social situation *around* the algorithmic system. Conceptualizing the situation as a game directs our attention to actions participants take in relation to their stakes in that situation (Lyman & Scott, 1989), along with the roles and relationships of the participants. In particular, we build on

[1]Aalto University, Finland
[2]University of Helsinki, Finland
[3]Stockholm University, Sweden

Corresponding Author:
Jesse Haapoja, Department of Computer Science, Aalto University, Konemiehentie 2, 02150 Espoo, Finland.
Email: jesse.haapoja@aalto.fi

Erving Goffman's work (1969) in which he studied human interactions with games as a heuristic device. Goffman's work points us to the roles adopted by actors involved in the game, their strategic moves, and their ways of controlling information in the situation. We use interviews conducted with members of the monitoring project and critical online discussions about the project to analyze how different stakeholders, with their differing interests, tried to achieve their conflicting goals and how the model was at the center of this conflict.

Our case is situated as part of a larger societal debate about the definition of hate speech, opinions regarding its harmfulness, and measures to mitigate it. Hateful, discriminating speech online, often targeted at minorities, has raised significant concerns (e.g., Gagliardone et al., 2015; Matamoros-Fernández, 2017), but the term 'hate speech' itself is broad and contested: Some definitions focus on hateful speech targeted at individuals or groups based on their ethnicity, gender, sexuality, or other sensitive personal properties. At the same time, hate speech is used "as a generic term, mixing concrete threats to individuals' and groups' security with cases in which people may be simply venting their anger against authority" (Gagliardone et al., 2015, p. 7). While definitional issues make it difficult to identify hate speech algorithmically, there are models that could help in monitoring it online (e.g., Burnap & Williams, 2015). However, even advanced hate-speech detection systems are argued to be vulnerable to deception (Gröndahl et al., 2018).

Platform companies try to counter problematic content with professional moderation (Gillespie, 2018), while online communities have deployed grassroots strategies, including Twitter blocklists (Geiger, 2016) and volunteer moderation (Matias, 2019). The project we explore resembles commercial content moderation (e.g., Gerrard, 2018; Gillespie, 2018) in terms of the logic of automation—both aim to differentiate the acceptable and the non-acceptable from online content—but here, the model was not operated by a platform company. Instead, the monitoring project was initiated by a non-governmental organization (NGO). Other participants included a governmental body tasked to prevent and tackle discrimination, a software company, another NGO, and a team of academic researchers. Candidates' Twitter and Facebook handles were collected, when available, from a website where they had answered to Voting Assistant Application questions and where they had had the possibility to fill in their campaigns' social media information. Facebook and Twitter APIs were used to query new posts by the candidates' public accounts daily. Data were collected from approximately 6,400 Facebook pages and 1,300 Twitter profiles as not all candidates had these or listed them in the Voting Assistant Application. The filtering of candidate posts was done using the model created by the software company. Researchers contributed by taking part in tagging the training data for the model to train it to detect hate speech (for more technical details, see Laaksonen et al., 2020).

Guided by Goffman's notion of games, we set out to explore how the game unfolded around the algorithmic model; What kinds of parties and roles emerged during the game? What types of moves they played? We examine different questions related to the game: why it exists, what are the payoffs, and how it is played on different levels. By levels, we refer to interactions between those who created the model and those critical of it, and between the model and those who oppose it. With this approach, we aim to understand a situation that features multiple stakeholders, conflicting aims, as well as human and non-human participants. Furthermore, on the conceptual level, we explore how broadening the metaphor of game beyond "gaming the system" might yield new insights regarding how people act in relation to algorithmic systems.

In what follows, we first discuss games as an analytical lens as well as prior research on strategic action with algorithms. We, then, present our materials and methods before moving on to findings, where we focus on the roles, stakes, and moves in the game. Our study illustrates how people not only play against algorithmic systems but also use them to play games with each other, creating and casting the system as an ally or framing it as an enemy. These systems, then, can be considered to be not only players, but also game pieces. Thus, algorithmic systems are mediators in a larger societal game played by humans, where they are mobilized to achieve certain goals and subsequently encounter resistance from other individuals. Building on this notion, we conclude by reflecting on the broader usefulness of the game metaphor in analyzing interactions with and around algorithmic systems.

'Game' as an Analytical Lens to Study Strategic Action With Algorithms

Several recent studies highlight user agency in relation to algorithms, often by discussing strategic actions that people engage in when encountering algorithmic systems. Most common examples come from social media platforms (e.g., Bucher, 2018; Cotter, 2018; Van Der Nagel, 2018; Witzenberger, 2018), search engines (e.g., Gillespie, 2017; Ziewitz, 2019) and gig work platforms (e.g., Chan, 2019; Lee et al., 2015). While working in the same stream, we approach automatic hate-speech detection and strategies to counter it through the metaphor of *a game*. A game as an analytical framework provides a way of ordering action and considering what participants have to lose or gain in a situation (Lyman & Scott, 1989). In other words, a game can be seen to begin when at least one actor thinks they have a stake in the situation. With this framing, not acting becomes a move in the game, too. Thus, our definition of *gaming* refers not only to moments where actors try to cheat or otherwise manipulate technical systems, but also to instances of interaction where something is at stake for someone. This is a broader definition of a game than the popular notion of 'gaming the system' (e.g., Bambauer & Zarsky, 2018).

We argue that interactions can almost always be approached with games as heuristic devices. We draw especially on Erving Goffman's (1969) scholarship on games. While Goffman (1969) is perhaps best known for his dramaturgical metaphor of everyday life, his book Strategic interaction, written partly while visiting game theorist Thomas Schelling in Harvard (Manning, 1992), considers the game-like elements of everyday life. We build on Goffman's work given its focus on different roles, moves, and the importance of controlling information in games, along with his considerations about the consequentiality of games (Goffman, 1967, 1969). In brief, Goffman distinguishes the roles of *party* and *player* in games—party referring to those whose interests are relevant for the game, while players are those who *act* on these interests. Although the same actor may be both a player and a party, this is not always the case—people sometimes act on others' interest, or negotiate common interests upon which to act on. In addition, Goffman considers how actors may be *tokens* (marking taken positions), or *informants* (who provide information to other actors in games). As for moves, Goffman discusses how actions of those involved in game-like situations can be classified based on whether they actively try to reveal some information about others or rather to dupe them. This control of information happens in what Goffman calls *expression games*. These are situations that deal with "the individual's capacity to acquire, reveal, and conceal information" (Goffman, 1969, p. 4), something that can be seen to be an element of all games. We use two particular categories of moves to discussed by Goffman (1969): control moves, which refer to "the intentional effort of an informant to produce expressions that he thinks will improve his situation if they are gleaned by the observer" (p. 12) and naïve moves, which refer to a situation where an observer thinks that the observed is acting naturally (p. 11).

Our work can be seen as a continuation of a long scholarly tradition that uses the concept of a game to study social life (Swedberg, 2001). At the same time, it parallels recent studies on algorithmic systems with a focus on the strategic behavior of individuals and groups, not only with the notion of a game (e.g., Aspling & Juhlin, 2017; Bambauer & Zarsky, 2018; Chan, 2019; Cotter, 2018; Ziewitz, 2019), but also with the help of de Certeau's (1988) concepts of tactics and strategies (Van Der Nagel, 2018; Willson, 2017; Witzenberger, 2018) and the lens of activism (Velkova & Kaun, 2019). Moreover, in the machine learning community, strategic manipulation has been used to describe situations where people try to influence classification systems by manipulating inputs into the systems (Hardt et al., 2016).

Much of this recent scholarship shares the idea that controlling who sees the information that participants reveal is consequential. These studies provide examples of the setting Crawford (2016) describes: "the spaces of intersection between humans and algorithms can be competitive and rivalrous, rather than being purely dictated by algorithms that are divorced from their human creators." For example, Cotter's (2018) article about Instagram influencers uses the game metaphor to study how influencers try to maximize their overall visibility, using different strategies to attain information about the platform's algorithm(s) and how they take advantage of such knowledge. Maximizing one's visibility includes strategies that fall into what Gillespie (2017) considers making one's actions algorithmically recognizable, that is, trying to make one's content more relevant by taking advantage of the logic of the algorithm. Users sometimes wish to manage their visibility with the aim of making it harder for some audiences to see or otherwise benefit from the information they share online by making it less algorithmically recognizable (Van der Nagel, 2018). Such action, however, requires awareness of the algorithms in question, that is, algorithmic imaginary (Bucher, 2017).

Materials and Methods

In line with Seaver's (2017) suggestion, we use diverse data sources to study an algorithmic system. One of the authors participated in the project from its beginning and also took part in tagging messages for hate speech to provide the model with training data. Moreover, the first author interviewed three participants from the project, each representing a different participating organization. Two of the interviewees represented different NGOs participating in the project, including the initiator (NGO1 and NGO2 in the results). The third interviewee was a representative of the participating governmental body (GOV in the results). While we were, unfortunately, unable to secure an interview from the software company that was involved, all relevant organizations are represented in our material, through the interviews and one author's direct participation in the project. All interviews were roughly an hour long, conducted in Finnish, and transcribed verbatim (resulting in 46 pages of transcription). The interviews were structured around four themes: (1) How the project started and how different actors joined it, (2) how it was run (division of tasks, challenges encountered), (3) the model itself (its creation, use, and consequences), and (4) what kind of aims, hopes, and expectations interviewees had had for the project and how they saw its outcomes.

To investigate opposing views, we analyzed two online forum threads that focused on the model, including altogether 230 messages. These discussions took place on a Finnish online forum popular among individuals who call themselves "immigration critics." This data source was chosen using the logic of purposive sampling (Silverman, 2006), that is, focusing on sources where processes of interest are most likely to be observed. Exploring recent content on the site, we first identified one discussion that directly dealt with the model. Later on, another discussion emerged that had messages concerning the model. Here, we focused on different strategies for countering the model or acting against the team behind it. These materials allowed us to study how the

model was authorized to act as a player by the team who created it, how it was received by others with contrasting viewpoints, and how the game itself was constructed by the participants.

In addition to interviews and forum discussions, we also draw upon informal discussions with the project team and news articles about the project. These serve us in gaining a broader understanding of the project and its consequences. For example, we heard through informal discussions that one project member had received an anonymous hostile message after being interviewed for a newspaper. News articles about the project were referred to in the online discussions, too, so we read the linked articles to better understand the discussions. We used these supplementary materials to help us make sense of the primary materials.

In our analysis of the primary materials, we draw on Goffman's (1967, 1969) work regarding games as a conceptual model. The first author used the concept of a game to guide the iterative analysis process. Originally, the analysis focused on expression games (Goffman, 1969), coding for (1) what kind of information different stakeholders tried to gain and (2) what kind of information they tried to signal. Here, we intended to focus on the "gaming the system" perspective, that is, on speculations and strategies about how to manipulate the model. However, the initial round of analysis revealed that this focus led to a limited understanding of the situation. Hence, the data were re-analyzed with codes relating to different roles, interests, stakes, and moves. As a result, we classified the *parties* and *players* in the game and identified different interests that were said to be served. Second, we explored how different participants in the game tried to make it matter to others, that is, what was at *stake* in the game. Here, a key idea was to consider what it would mean for someone to win or lose in the game. Third, we examined *expression games* with the model: how was the model used to try to reveal information, and what kinds of strategies did those opposing the model discuss for manipulating or hiding information from it? Here, we drew on our different materials to understand in more detail the model's role in the larger game around it. Finally, we revisited the transcripts to ensure that our analysis reflects the materials fairly and does not include exaggerated claims.

Parties and Players in the Game

When looking at algorithmic systems through the theoretical lens of a game as we have conceptualized it here, we consider the users, system designers, and other stakeholders as *parties* who have interests regarding the outcome of the game and/or as *players* who act on those interests (Goffman, 1969). In addition, we conceptualize the game that was played with and around the model as one where its designers and those critical of it could be both *parties* and *players*. As mentioned above, the difference between these two is that a party has a unitary interest to promote while a player is authorized to act on a party's behalf (Goffman, 1969, p. 86). A member of a party can be a player, but these roles are not always shared by the same actor. As algorithms do not have interests, they may be players but not parties— in this case, their role is to represent the interests of others. To be clear, a party in this sense is not the same as *a political party*. Below, we describe how roles in the game became visible. Examining how different parties and players position themselves in relation to each other helps us to understand why a game exists and what it is about. When it comes to the empirical case at hand, this helps us make sense of questions like why particular players played (or speculated on playing) the moves we describe below, how they formed parties under common interests, and how the roles they found themselves in were brought about in their relations to the other party and players.

The game we describe here came into being through the following process: first, a coalition of organizations crafted a common goal of using machine learning techniques to detect hate speech in municipal elections. Further individuals who heard about this intervention and considered it problematic then discussed strategies to counter the effort. Here, we frame this disruption of routine as the beginning of a game that formed around the unitary interests shared within the two parties and the conflicting interests between them. As existing literature points out, the same technology can be interpreted in varied ways (Orlikowski & Gash, 1994; Raita, 2012) and can be seen as either useful or as a hindrance, depending on the position it is interpreted from (Raita, 2012). In our analysis, the division could be said to be that of an ally or an enemy: one interviewee even stated that she was interested in finding out "what the machine is capable of as a friend" (NGO2). In what follows, we approach the game around the model from the perspective of it either as an allied player (for those developing it) or as an enemy (for those opposed to it).

Creating an Allied Player

There was a consensus among our interviewees that with social media, new forms of hate speech have emerged (one interviewee, however, was critical about the idea that social media has increased hate speech, since it is hard to prove such a claim.) Those interviewed also shared a hope that the model would provide benefits of scale by being able to go through messages more efficiently than a human. In all interviews, hate speech was seen as a problem, and there was willingness to test new ways of countering it on social media. The party advancing the project was formed around this idea and shared ambition.

Despite the shared ambitions, the representative of the governmental organization stated that she felt there were differing goals among project participants. For example, she thought that for the software company, testing and building the model for technical learning purposes was perhaps a

more central interest than the hate speech monitoring aim itself, while their own goal was fairly clear:

> We had different goals regarding what we wanted to do, but we wanted to preempt hate speech in municipal elections through political parties and affect more in that way. The whole technical implementation and algorithm were more of a side project for us but we were interested to try what could be achieved. (GOV)

She also stated that "reducing the amount of hate speech is quite an obvious goal for a public organization whose mission is to improve equality" (GOV). So, their organization had a clear reason for participating in the project, and while it might have differed from those of others, there was enough common ground to work together.

Our interviewees brought up multiple times that all political parties in the parliament have signed a treaty against racism, including hateful speech concerning ethnic groups. In the discourse, they reason the treaty as a way for those involved in the project to position themselves so that their use of the model as a player was justified. In the words of one interviewee, "We monitor if the parties are following the rules which they have stated that they will follow" (NGO1). The interviewee from another NGO argued that the model only went through politicians' public messages and, for that reason, it had the right to "read" them. Since election candidates are public actors, the team behind the model deemed it acceptable to supervise their social media activities, whereas the same would not be true for the general public who has rights to be online without such monitoring: "We discussed the limits so that we do not breach privacy. . . all candidates publish information by themselves, so they are public in that sense and they have a position as a candidate for a political party" (NGO1). In other words, candidates' right to privacy was lowered due to their public role. Moreover, the analyzed messages were publicly available.

The creation of the model constructed potential opposing players to those involved in the project: candidates who use or want to use expressions that fall under the definition of hate speech in social media during their campaigning (or feel that others should be allowed to do so). Those who created and deployed the model were not able to be certain that there would be any candidates who would act in a way that would get their messages tagged by the model.

Reacting to an Enemy Player

Not everyone agreed with the goals that the system was built to achieve. In the online forum data we collected, those critical of the project framed their criticism by claiming that the system was detrimental to the freedom of speech and some even characterized the act of monitoring as "Orwellian." The opposing party positioned itself as protecting the freedom of speech while claiming that those creating the system oppose—or even actively counteract it: "So this is how the freedom of speech is destroyed. Wrong-thinkers are silenced while good people cheer. Humanity has not learned anything." This stance was also used to argue that the party behind the system did not have the right to deploy the system. In particular, those critical of the project stated that the governmental organization that took part in the project had overstepped its mandate. The online discussions we analyzed implied that the opposing party saw the governmental organization as a long-standing enemy and that the animosity visible in the material was not particular to only this occasion. The governmental body representative interviewed also stated that there is a group that actively criticizes them whenever they work on related issues publicly. To sum, this party considered the model to represent their enemies and, thus, the model became something that needed to be countered.

It should be noted here that while some discussants in our forum data did state that they were candidates in the election, it should be expected that most of them were not running for office themselves. Even so, the discussants tended to align themselves with those candidates—imagined or real—that the model was seen to threaten. Thus, there was a sense of the party having a shared interest to protect. At the same time, the party constructed those who developed and deployed the model as an opposing party. Through these steps, the situation got framed as an instance of a classical game of "us versus them."

Stakes of the Game

To better understand the game in question, it is important to have a sense of what the different parties felt was at stake and what consequences the game might have for them. In this section, we discuss how the game was made to matter for those involved. We conceptualize actions directed to cause consequences for the opposing party as *moves* in the game, that is, courses of action that involve real consequences in the external world and that alter the situation of the game's participants (Goffman, 1969, p. 90).

Making the Game Matter for Those Involved

The messages that the model tagged as potential hate speech lead to subsequent moves by representatives of the governmental body. They had planned to inform the police about messages that broke the law, but none of the messages tagged was deemed severe enough for this. When it came to content that was not illegal but that was clearly a violation of the anti-racist treaty signed by all parliament parties, the authorities contacted the political party of the candidate to inform them about the issue.[1] We can consider consequences like legal action or being reprimanded (or worse) by one's political party as potential negative consequences which made the game matter for its players.

Similarly, those who resisted the model also planned or played out moves against the deployers of the model. The act

of participating in the project publicly can itself be seen as a potentially risky move. It is not uncommon that right-wing actors start to ridicule or even harass those that publicly act against racism or hate speech (e.g., Hatakka, 2019). Criticism toward the people behind the system was common in the online discussions analyzed for this study, too.[2] Some of the project participants accounted that they had received hostile messages. Some forum messages included names of individuals from participating organizations, shared with the intention to expose these individuals to scrutiny by those reading the forum discussion (as is common in acts of "doxing," for example, Douglas, 2016). An especially noteworthy example was a message where a discussant not only listed all board members of one participating NGO (instead of only those publicly participating in the project) but also added links to their LinkedIn profiles, stating that "the NGO's other board members are all in LinkedIn. You can go there and read about their adventures in academia from one sinecure to another."

While the party behind the model emphasized how they made sure no laws were broken and, thus, had nothing to worry about from that perspective, there was some discussion in our online forum data advocating for the illegality of the project. Claims regarding the (il)legality of the project were made based on freely available information, namely, news articles written about the project. In an attempt to delegitimize the party they felt was acting against them, some discussants referred the Finnish Personal Data Act and tried to argue how it had been broken: "Law does not identify hate speech, and making illegal registries is illegal." Others argued that the governmental body had exceeded their authority.

A concrete move based on this line of reasoning was played when one member of the parliament delivered an interpellation about the system and its use to the minister of justice of the Finnish government, as the governmental organization participating in the project falls under this minister's jurisdiction. Two forum discussants had stated that some like-minded parliament member should do this before it happened. These actions can be seen as moves where the attempt is to uncover some slip-up that would give a reason for law enforcement to act against the creators of the model, and thus, cause negative payoffs to them.

An Expression Game With the Model

We now turn to consider how the model and its opponents face each other, building on the concept of *expression games*. This concept focuses on how, in games, there are players that are observers and/or observed. According to Goffman (1969), expression games are part "of something more inclusive, a game concerning objective courses of action" (p. 145). So far, we have focused on how information is used for attempts to punish opponents. We now consider how the model was used as a part of an expression game, where to goal set for the model was not only to collect information and thus act as *an informant* (Goffman, 1969, p. 88) but also to signal that the candidates were being monitored, and so the model acts as *a token* (Goffman, 1969, p. 87). In addition, we analyze how those opposing the model devised strategies to counter it.

Model Deployment as a Control Move

Deploying the model as a player can be seen as a control move. The control moves of interest here are acts to "reveal as unmistakably as possible" (Goffman, 1969, p. 17) what was done. As a *token*, the model was part of a larger project to preempt or discourage hate speech. Its capability to be significant as a token relied on its other role as an *informant.* The act of employing the algorithm can be seen as a move that was meant to be noticed. This was achieved with the help of press releases authored by the participating organizations and picked up in media reporting about the project.

Moving the algorithm into the field was an action that was supposed to be taken seriously: one interviewee stated that this showed that "it is not only the police that is watching the actions of the candidates" (NGO2). In other words, the model's role as a token was to claim a presence in social media for the party behind it and to show that the candidates are being monitored. In a remark that implicitly acknowledges the power of this move, one online discussant speculated that the model might only exist as a scare: "I think that they could be able to go through the candidates' messages manually, maybe this opinion checking system has been created to scare the supporters." Deploying the model meant that the candidates were trapped in the game if they used public Facebook pages or Twitter accounts in their campaigning (given that they reported them in the Voting Assistant Application). Whether they wanted it or not, if they were active on public social media feeds, their messages were monitored by the model. As such, publicly moving the model to "the game board" held power as it was used to change the definition of the situation so that publishing social media messages that might be classified as hate speech became riskier for the candidates. This was in line with the governmental organization's goals: "We have made the decision that we want to influence those that hold great power in the society, that is, the decision-makers" (GOV), referring not only to this project but their overall aim in regard to hate speech. Moreover, given the model's role as an informant—a player that monitored the candidates' actions—even those unaware of the project became, unknowingly, players in the game. The model's deployment was a combination of two moves: on the one hand, there were wishes to identify those who used hate speech; on the other, the model was simultaneously hoped to mitigate toxic language used by the candidates.

Model's Actions as Naïve Moves

When the model is considered as an informant, its actions in the field can be seen as naïve moves, playing against those who are not necessarily aware that their actions are being monitored. An interviewee from one of the participating organizations hoped that the model could reveal actions made by less-known candidates that might otherwise slip through unnoticed. Another interviewee described municipal elections as a good event to test the model since the breadth of candidates was likely to mean that, on the whole, they would use more "authentic language" than professional politicians:

> There may still be quite a lot of candidates in the municipal elections who have not necessarily honed their language to match certain standards of conduct, so that way there is a possibility to capture authentic language and sadly, authentic hate speech or racism. (NGO1)

By this, he referred to an idea that in parliament or presidential elections, candidates are already so accustomed to political life that, even if they harbor prejudiced thoughts, they are likely to articulate them in a way that does not breach any laws or established ideas of correct conduct. Thus, more experienced politicians would know how to "play the game."

Finally, we need to consider the possibility of individuals attempting control moves that aim at fooling the model. Since algorithms cannot be reflexive (Alkhatib & Bernstein, 2019), they cannot consider that they might be duped. Their rigid and consistent nature makes them vulnerable for manipulation attempts (Bambauer & Zarsky, 2018). The model is naïve even though the individuals behind it are aware that it could be fooled. An NGO representative stated that "some circumlocutions could be used so that it (a message containing hate speech) would go unnoticed" (NGO1), referring to the possibility that individuals might use language in creative ways to counter the model.

Control Moves by the Party Opposing the Model

The act of deploying the model transformed candidates' situation in a way that prompted those critical of the system to consider how to counter it. Potential control moves were discussed on the online forum, in a process what Bishop (2019) refers to as *algorithmic gossip*. Here, the party formed theories of the model socially, through discussion with other members. These were based on the idea that the model could only embark on naïve moves, that is, the model could not take into account the possibility of it being fooled. This can be understood, in Goffman's (1969) terms, as speculation about the opponent's style of play and its potential moves (p. 95). Such speculation was evident in messages that pondered on what kind of word lists the model uses. Technically, this was misguided, since the model was not built with any explicit word lists gathered by the team behind it. Instead, the model relied on classical supervised machine learning techniques and was trained with example messages that were considered hate speech and others that were not. Yet, the speculation did reveal ideas regarding the kind of content participants on this side of the game thought to be considered relevant by the opposing team. Moreover, a few discussants reflected on the model based on their knowledge about machine learning methods. For example, one discussant stated that

> This could be achieved with machine learning, where you need to feed the algorithm lots of teaching data where every message has to be tagged by whether it has "hate speech" in it or not. This requires a lot of data.

and continued that "Coding it does not seem like rocket science. The problem will be to create a completely intelligent identification algorithm. This is what even Facebook has problems with." This type of discussion was in line with how the project was actually carried out. Prior experiences were used to make sense of the opposing player, its way of playing, and its weaknesses. These excerpts also reveal different algorithmic imaginaries (Bucher, 2017) that individuals taking part in the discussion had about the model.

One of the strategies discussed was obfuscation (Brunton & Nissenbaum, 2015). The idea here was to use words or entire messages that the discussants thought would end up filtered in by the model but in a way that would not be punishable. Using Gillespie's term (2017), content that the team behind the system would not like to receive would be made *algorithmically recognizable*, that is, efforts would be made to make it seem relevant to the model. There would, then, be too much content for the opposing party to be able to act upon. One example of this sort of strategy shows how prior encounters with moderation are used in a new context:

> Old trick that was used in Usenet's news could be effective. You could add to the end of every Facebook-post a sentence: three randomly chosen words from the dictionary: asylum muslim terrorist. Then you can just change the words every now and then. After this, almost every single message will be filtered in and the censors will drown in the flood of messages since every single message has to be gone through manually.

These types of strategies drew from the interpretation that the model cannot distinguish between the meanings that different words have in different contexts. Further ideation along this line leads some discussants to consider a script that would generate lists of words automatically—using algorithms to game an algorithm (the model).

Another strand of potential control moves were those where the message could be understood by humans but not necessarily by the model. One discussant pondered whether

the model could be fooled by splitting words from the middle. It was also speculated that the fairly complicated nature of Finnish language would be too hard for the model to filter. This speculation falls into line with the Van der Nagel's (2018) insights regarding how content can be made difficult for machines to understand while keeping it easy for humans to decipher.

These strategies reveal how the model and the party that deployed it were constructed by those discussing it. First, the model may have some weaknesses that the party behind it was either unable to solve or did not take into consideration. Second, the party itself has the weakness of having limited time and resources to spare. The governmental representative stated that the system filtered too many messages for manual checking, making it less valuable for them. From this perspective, it could be argued that the weaknesses discussed were accurate. On the contrary, models can be trained further. While the model itself cannot understand if it is being duped, its designers can, and they may react accordingly or even counter preemptively the more obvious exploits (Haapoja & Lampinen, 2018).

Overall, some of the strategies discussed would have been more viable if the model was targeted to a larger audience, but as it was only monitoring election candidates, it could have been detrimental for their campaigning had they started to use their accounts to game the system. However, this is a further piece of evidence that individuals speculate about the functioning of algorithms and may embark on strategies to cope with or counter them, as discussed in prior studies (e.g., Bishop, 2019; Bucher, 2017, 2018; Cotter, 2018; Lee et al., 2015; Rosenblat & Stark, 2016; Van der Nagel, 2018).

Discussion and Conclusion

While considering social life as a game is an age-old trick, our daily encounters with algorithmic systems can be seen to give rise to new types of games, where alliances and antagonisms are formed, expressed, and played out in technologically mediated environments. Using a game as a metaphor, we interpreted our case as one where a set of actors formed an alliance to counter hate speech with the help of machine learning, and another group reacted to this negatively, trying subsequently to devise counter strategies. In other words, while one party saw the technology as a potential ally, another rejected it as an enemy, questioning its right to even exist. While the model itself was an important part of the game, we argue that a sole focus on interactions *with* the model would have led to a limited understanding of our case where significant interactions took place also *around* the model.

We used the game metaphor to highlight individuals as active agents with their own goals, interpretations, and preferences when creating and encountering algorithmic systems. This aligns with Cotter's (2018) call to treat the targets of algorithmic techniques as agents who act strategically to achieve their goals rather than as passive cogs in the system. Our study cannot (and is not meant to) represent all encounters with algorithmic systems, but it complements prior literature with an empirical case that considers also those behind the model—an approach more difficult to adopt when scrutinizing global, commercial systems. Access to the model's creators allowed us to illuminate their agency, their goals, and the ways in which these were translated into the model when it was developed and deployed. The game metaphor served us in making sense of how different actors gather together as parties to create, use, and potentially resist algorithmic systems. It foregrounds different meanings that can be assigned to a system depending on how individuals and parties position themselves in relation to it. This ties strongly to questions of whose interest particular technologies (are seen to) advance.

In our study, we conceptualized the model as a token which, when deployed to the field, had the power to alter the meaning of social media activities in a way that made certain behaviors riskier for the candidates, rendering those deploying the system (and opposed to hate speech) more powerful. Systems that alter the status quo may disrupt routinized activities, for example, by enforcing norms in more effective ways, and these power shifts may even, at first, go unnoticed by some that are affected by them. This resonates with prior research on power asymmetries between different stakeholders in relation to algorithmic systems (Beer, 2017). For example, parties in charge of algorithms may delegate to them the capability to reward or punish certain actions. This is commonly seen in how social media platforms grant visibility to their users (Bucher, 2018; Cotter, 2018) and in how gig work platforms such as Uber control the workers affiliated with them (Chan, 2019; Lee et al., 2015).

As has been brought up in prior studies, too, algorithms' weaknesses as players are defined by their incapability for reflexivity (Alkhatib & Bernstein, 2019) and by human efforts to identify gaps in their capability to perceive and interpret action (Van der Nagel, 2018; Witzenberger, 2018). In our case, the model was seen as capable of learning rules but vulnerable in its inability to diverge from them. As pointed out previously (e.g., Bucher, 2018; Cotter, 2018; Van der Nagel, 2018), people can, even with limited knowledge about different systems, exercise agency in relation to them. Moreover, those creating algorithms can experiment with countering human manipulations. Manipulation by users and subsequent responses by designers can be understood as an ongoing game that renders visible the agency of humans on both sides of particular systems. Our case here was somewhat extreme in that the resisting of the deployed model was evident and observable. However, we propose that the game metaphor is a useful analytical tool also more broadly in studying the role of algorithmic systems in society. Algorithms are directly a part of the game on some levels, and have a less active role on others. As such, the game metaphor can be productively expanded beyond the conceptualization of gaming as a particular form of relating to the system by human users and

applied also to the broader, messier game surrounding these specific interactions.

Acknowledgements

The authors thank the anonymous reviewers, Marisa Cohn, Kari Vesala, Barry Brown and the participants of Rajapinta-skrivarstuga and the NOS-HS Nordic Perspectives on Algorithmic Systems workshop series for their invaluable comments on different versions of this manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the Kone Foundation project Algorithmic Systems, Power, and Interaction, the Academy of Finland grant 295948 and the Swedish Foundation for Strategic Research project RIT15-0046.

ORCID iDs

Jesse Haapoja https://orcid.org/0000-0001-6877-7957
Salla-Maaria Laaksonen https://orcid.org/0000-0003-3532-2387
Airi Lampinen https://orcid.org/0000-0002-9100-3826

Notes

1. Finnish law does not identify hate speech as a crime: juridical cases related to hate speech are based on laws about incitement of hatred or defamation (Pöyhtäri et al., 2013).
2. There were no direct threats or other messages that would account for legal action in the forum data.

References

Alkhatib, A., & Bernstein, M. (2019). Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery

Aspling, F., & Juhlin, O. (2017). Theorizing animal-computer interaction as machinations. *International Journal of Human-Computer Studies*, *98*, 135–149.

Bambauer, J., & Zarsky, T. (2018). The algorithm game. *Notre Dame Law Review*, *94*, Article 1.

Beer, D. (2017). The social power of algorithms. *Information Communication and Society*, *20*(1), 1–13.

Bishop, S. (2019). Managing visibility on YouTube through algorithmic gossip. *New Media & Society*, *21*(11–12), 2589–2606. https://doi.org/10.1177/1461444819854731

Brunton, F., & Nissenbaum, H. F. (2015). *Obfuscation: A user's guide for privacy and protest*. MIT Press.

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, *20*(1), 30–44.

Bucher, T. (2018). *If . . . then: Algorithmic power and politics*. Oxford University Press.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*(2), 223–242.

Chan, N. K. (2019). The rating game: The discipline of Uber's user-generated ratings. *Surveillance & Society*, *17*(1/2), 183–190.

Cotter, K. (2018). Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New Media & Society*, *21*(4), 895–913. https://doi.org/10.1177/1461444818815684

Crawford, K. (2016). Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Science, Technology, & Human Values*, *41*(1), 77–92.

De Certeau, M. (1988). *The practice of everyday life*. University of California Press.

Douglas, D. M. (2016). Doxing: A conceptual analysis. *Ethics and Information Technology*, *18*(3), 199–210.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech: UNESCO series of Internet freedom*. United Nations Educational, Scientific and Cultural Organization. http://unesdoc.unesco.org/images/0023/002332/233231e.pdf

Geiger, R. S. (2016). Bot-based collective blocklists in Twitter: The counter public moderation of harassment in a networked public space. *Information, Communication & Society*, *19*(6), 787–803.

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, *20*(12), 4492–4451.

Gillespie, T. (2017). Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. *Information, Communication and Society*, *20*(1), 63–80.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Goffman, E. (1959). *The presentation of self in everyday life*. Doubleday.

Goffman, E. (1967). *Interaction ritual: Essays on face-to-face interaction*. Pantheon Books.

Goffman, E. (1969). *Strategic interaction*. University of Pennsylvania Press.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is "Love." In *Proceedings of the 11th ACM workshop on artificial intelligence and security: AISec'18* (pp. 2–12). Association for Computing Machinery Press.

Haapoja, J., & Lampinen, A. (2018). "Datafied" reading: Framing behavioral data and algorithmic news recommendations. In *Proceedings of the 10th Nordic conference on human-computer interaction* (pp. 125–136). Association for Computing Machinery Press. https://dl.acm.org/doi/10.1145/3240167.3240194

Hardt, M., Megiddo, N., Papadimitriou, C., & Wootters, M. (2016). Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science: ITCS'16* (pp. 111–122). Association for Computing Machinery Press. https://dl.acm.org/doi/10.1145/2840728.2840730

Hatakka, N. (2019). Expose, debunk, ridicule, resist! Networked civic monitoring of populist radical right online action in Finland. *Information, Communication & Society*. Advance online publication. https://doi.org/10.1080/1369118X.2019.1566392

Laaksonen, S.-M., Haapoja, J., Kinnunen, T., Nelimarkka, M., & Pöyhtäri, R. (2020). The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring. *Frontiers in Big Data*, *3*, Article 3. https://doi.org/10.3389/fdata.2020.00003

Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. (2015). Working with machines. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems: CHI'15* (pp. 1603–1612). Association for Computing Machinery Press.

Lyman, S. M., & Scott, M. B. (1989). *A sociology of the absurd*. General Hall.

Manning, P. (1992). *Erving Goffman and modern sociology*. Polity Press.

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society*, *20*(6), 930–946.

Matias, J. N. (2019). The civic labor of volunteer moderators online. *Social Media + Society*, *5*(2), Article 983677. https://doi.org/10.1177/2056305119836778

Orlikowski, W. J., & Gash, D. C. (1994). Technological frames: Making sense of information technology in organizations. *ACM Transactions on Information Systems*, *12*(2), 174–207.

Pöyhtäri, R., Haara, P., & Raittila, P. (2013). *Vihapuhe sananvapautta kaventamassa*. Tampere University Press.

Raita, E. (2012). User interviews revisited: Identifying user positions and system interpretations. In *Proceedings of the 7th Nordic conference on human-computer interaction: Making sense* (pp. 675–682). Association for Computing Machinery Press. https://dl.acm.org/doi/10.1145/2399016.2399119

Rosenblat, A., & Stark, L. (2016). Algorithmic labor and information asymmetries: A case study of Uber's drivers. *International Journal of Communication*, *10*, Article 27.

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2), Article 738104.

Silverman, D. (2006). *Interpreting qualitative data: Methods for analyzing talk, text and interaction*. SAGE.

Swedberg, R. (2001). Sociology and game theory: Contemporary and historical perspectives. *Renewal and Critique in Social Theory*, *30*(3), 301–335.

Van Der Nagel, E. (2018). "Networks that work too well": Intervening in algorithmic connections. *Media International Australia*, *168*(1), 81–92.

Velkova, J., & Kaun, A. (2019). Algorithmic resistance: Media practices and the politics of repair. *Information, Communication & Society*. Advance online publication. https://doi.org/10.1080/1369118X.2019.1657162

Willson, M. (2017). Algorithms (and the) everyday. *Information, Communication & Society*, *20*(1), 137–150.

Witzenberger, K. (2018). The hyperdodge: How users resist algorithmic objects in everyday life. *Media Theory*, *2*(2), 29–51.

Ziewitz, M. (2019). Rethinking gaming: The ethical work of optimization in web search engines. *Social Studies of Science*, *49*(5), 707–731. https://doi.org/10.1177/0306312719865607

Author Biographies

Jesse Haapoja (M.Soc.Sci., University of Helsinki) is a PhD Student in Social Psychology at the University of Helsinki and a researcher at the Department of Computer Science at Aalto University. His research interests include micro-sociological and relational approaches to algorithmic systems and human–computer interaction.

Salla-Maaria Laaksonen (Dr. Soc.Sc., University of Helsinki) is a postdoctoral researcher at the Centre for Consumer Society Research at the University of Helsinki. Her research interests include the use the data and algorithms in organizations and the ways in which platform technologies are changing our public communication flows.

Airi Lampinen (Dr. Soc.Sc., University of Helsinki) is an associate professor of Human–Computer Interaction at the Computer and Systems Sciences Department at Stockholm University and a Docent of Social Psychology at the University of Helsinki. Her research interests include networked platforms, algorithmic systems, and the interweaving of interpersonal and economic encounters.